Accounting for Human Engagement Behavior to Enhance AI-Assisted Decision Making

Ming Yin

Purdue University, USA
mingyin@purdue.edu

Abstract

Artificial intelligence (AI) technologies have been increasingly integrated into human workflows. For example, the usage of AI-based decision aids in human decision-making processes has resulted in a new paradigm of *AI-assisted decision making*—that is, the AI-based decision aid provides a decision recommendation to the human decision makers, while humans make the final decision. The increasing prevalence of human-AI collaborative decision making highlights the need to understand how humans engage with the AI-based decision aid in these decision-making processes, and how to promote the effectiveness of the human-AI team in decision making. In this talk, I'll discuss a few examples illustrating that when AI is used to assist humans—both an individual decision maker or a group of decision makers—in decision making, people's engagement with the AI assistance is largely subject to their heuristics and biases, rather than careful deliberation of the respective strengths and limitations of AI and themselves. I'll then describe how to enhance AI-assisted decision making by accounting for human engagement behavior in the designs of AI-based decision aids. For example, AI recommendations can be presented to decision makers in a way that promotes their appropriate trust and reliance on AI by leveraging or mitigating human biases, informed by the analysis of human competence in decision making. Alternatively, AI-assisted decision making can be improved by developing AI models that can anticipate and adapt to the engagement behavior of human decision makers.

Introduction

The rapid development of artificial intelligence (AI) technologies has made profound impact on the human society in the past decade. For example, the rise of AI-based decision aids has fundamentally transformed how decisions are made. In the past, given a decision making task, it is usually humans who will process the task, deliberate about different decision options, and make the decision. Today, as AI models are developed to uncover hidden insights from the big data, given a decision making task, decision aids powered by AI models can provide decision recommendations, which will then be presented to human decision makers to assist them in their final decision making. This new paradigm of "*AI-assisted decision making*" implies that decision making nowadays is a collaborative task that is jointly completed by humans and AI. The promise is that with the human judgment and contextual awareness coupled with the analytical prowess and data-driven insights of AI, humans and AI can complement each other and achieve a level of decision making performance that exceeds the performance of either party alone.

In practice, however, it is found that such "*human-AI complementarity*" is rarely achieved (Hemmer et al. 2021; Tan et al. 2018). Our research suggests that a fundamental reason behind this unsatisfactory human-AI team performance in AI-assisted decision making is *a lack of sufficient and effective engagement* with AI among the human decision makers—instead of carefully analyzing the respective strengths and limitations of AI and themselves, human decision makers tend to engage with AI-based decision aids in a heuristic and biased way that results in uncalibrated trust and reliance on AI. As such, enhancing AI-assisted decision making requires a fundamental rethinking of the designs of AI-based decision aids. That is, instead of simply providing the most "accurate" decision recommendations, AI-based decision aids should be designed to account for humans' engagement behavior and strive to cultivate and optimize the engagement of human decision makers from the outset. This goal can be achieved by both introducing interventions in AI-based decision aids to influence humans' engagement behavior, and incorporating considerations of humans' engagement behavior in the training objectives of the AI models underlying the decision aids.

In the following, we will provide examples illustrating how humans engage with AI-based decision aids in practice, and how to account for human engagement behavior in the designs of AI-based decision aids to improve the human-AI team performance when AI supports either individuals or groups in their decision making.

Accounting for Engagement Behavior When AI Assists Individuals in Decision Making

Understanding human engagement behavior. Previous research has shown that when AI-based decision aids assist an individual decision maker's decision making, the individual's trust in the AI model's decision recommen-

dations as well as how much they are willing to rely on these recommendations are largely shaped by the AI model's performance (Yin, Wortman Vaughan, and Wallach 2019; Rechkemmer and Yin 2022). However, in practice, there are many scenarios that humans need to engage with the AI model's decision recommendations without knowing any performance information about the AI model. How will humans decide how much to trust and rely on the AI recommendations in this case?

One of our key observations is that in such scenarios, humans often leverage their confidence in their own independent judgement to decide how to engage with AI. This is because they tend to believe they are correct when they are confident (despite this is not always true as humans may sometimes suffer from the "*Dunning-Kruger effect*"). More specifically, it was found that when the performance information about the AI model is absent, humans are more likely to rely on their own judgement when their own decision confidence is high, while they are more receptive to AI recommendation when their own decision confidence is low (Wang, Lu, and Yin 2022; Chong et al. 2022). In addition, the level of *agreement* between the AI recommendation and humans' independent judgement on those decision making tasks where humans are highly confident about their own judgement also significantly impacts humans' trust and reliance on AI—the higher the level of agreement, the more humans trust and rely on the AI recommendation (Lu and Yin 2021). In other words, humans tend to use the level of high confidence human-AI agreement as a heuristic to gauge the trustworthiness of the AI model, which may, to some degree, reflect humans' "*confirmation bias*" towards AI. Such biased engagement behavior may result in humans' over-reliance (or under-reliance) on AI simply because AI tends to agree (or disagree) with humans. Even worse, such bias may lead to the lowest level of human-AI team performance in decision making when humans are assisted by the most "complementary" AI model (since complementary AI excels at different tasks than humans and therefore tends to disagree with humans).

Designing AI-based decision aids to account for human engagement behavior. In light of our understandings of individual decision makers' biased engagement with AI-based decision aids, we next explore how to design AI-based decision aids to account for the human engagement behavior. One possibility is to facilitate humans to engage with AI more "rationally" through analyzing humans' true competence on different decision making tasks, especially in relative to AI (Ma et al. 2023). Specifically, given a decision making task, we may estimate the correctness likelihood of both humans' independent judgement and the AI recommendation on this task. Depending on the correctness likelihood comparison of humans and AI, we can dynamically change how AI recommendation is presented to humans to nudge humans to rely more on the party who has the higher correctness likelihood. For example, when the human's correctness likelihood is estimated to be lower than that of the AI model's on a task, we can directly present the AI recommendation to the human decision maker upfront. Otherwise, we can ask the human decision maker to make an independent judgement before seeing the AI recommendation—this effective serves as a "cognitive forcing function", which has previously been shown to decrease humans' reliance on AI recommendations (Buçinca, Malaya, and Gajos 2021).

A complementary approach is to accept humans' biased engagement behavior as is, and incorporate such human behavior into the AI training process to obtain "behavior-aware" AI models that can directly optimize for the human-AI team performance. For example, as humans are found to rely on AI recommendation more when their self-confidence is low, an intuitive idea is that the AI model can be trained to excel on those tasks where humans have low confidence themselves and therefore "need" accurate AI recommendation more. We have shown that under a threshold-based model characterizing when humans will adopt the AI recommendation, training an AI model to optimize for the human-AI team performance in decision making effectively becomes solving a weighted empirical risk minimization problem, where the weight associated with each training instance is inversely proportional to the human decision maker's self-confidence on it (Mahmood, Lu, and Yin 2023). Human-subject studies further confirm that when humans are assisted by a behavior-aware AI model, they can achieve a significantly higher level of decision making accuracy than those who are assisted by the standard AI model.

Accounting for Engagement Behavior When AI Assists Groups in Decision Making

Understanding human engagement behavior. In real life, many decisions are made by a group of decision makers rather than an individual decision maker, and AI-based decision aids can also be used to support a group of decision makers in their decision making. A natural question of interests is how a group of decision makers may engage with AI-based decision aids similarly or differently than an individual decision maker. Our comparative study (Chiang et al. 2023) shows that compared to individuals, groups exhibit higher levels of reliance on AI recommendations in general, which leads to both a lower level of under-reliance on AI and a higher level of over-reliance on AI. Qualitative analysis of groups' discussion log further reveals a few possible reasons underlying this increased level of reliance on AI. For example, some group members may use the fact that the AI recommendation agrees with their judgement as a way to convince other members in their group, while other groups may use the AI recommendation as a tiebreaker when they can not reach a consensus. Further, groups' tendency to rely on AI may be amplified by the "*Bandwagon effect*", that is, people may easily agree to adopt AI recommendation when others are doing so (or suggest to do so).

Designing AI-based decision aids to account for human engagement behavior. As groups tend to over-rely on AI recommendations, enhancing AI-assisted group decision making requires us to design AI-based decision aids to encourage groups' critical reflection of the AI recommendation's trustworthiness. To this end, one intuitive idea is to introduce a "devil's advocate" in group discussion during the

decision-making process to force the group engaging in thorough deliberation. Traditionally, humans are asked to play the role of devil's advocate to present the dissenting argument, which may not be the most powerful due to a lack of "authenticity", and humans may even experience a threat to their psychological safety by playing this role. To overcome this limitation, we propose to incorporate a devil's advocate that is powered by large language models (LLM) into AI-assisted group decision making (Chiang et al. 2024). Our experimental results show that when LLMs are instructed to argue against the AI recommendation, and when they can actively participate in the group discussion in response to arguments made by other members in the group, the presence of these LLM-powered devil's advocate can help groups rely on AI-based decision aids more appropriately, and result in higher levels of decision accuracy.

Conclusion

The human-AI team performance in AI-assisted decision making is not only decided by humans' and AI's independent decision making performance alone, but is also largely shaped by humans' engagement with AI-based decision aids, e.g., how humans factor AI recommendations into their final decision making. Enhancing AI-assisted decision making—both when AI assists an individual decision maker and when AI assists a group of decision makers—requires us to first obtain a thorough understanding of how humans engage with AI in decision making. Our research suggests that humans often engage with AI-based decision aids in a heuristic and biased manner, which inspires us to re-examine the designs of AI-based decision aids to account for these engagement behavior. We find that AI-assisted decision making performance can be improved when the decision aids are explicitly designed to cultivate and optimize the engagement from human decision makers, by including interventions to promote appropriate engagement behavior or incorporating human behavior into the AI model's training objectives to directly optimize for the human-AI team performance.

References

Buçinca, Z.; Malaya, M. B.; and Gajos, K. Z. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1): 1–21.

Chiang, C.-W.; Lu, Z.; Li, Z.; and Yin, M. 2023. Are Two Heads Better Than One in AI-Assisted Decision Making? Comparing the Behavior and Performance of Groups and Individuals in Human-AI Collaborative Recidivism Risk Assessment. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–18.

Chiang, C.-W.; Lu, Z.; Li, Z.; and Yin, M. 2024. Enhancing AI-Assisted Group Decision Making through LLM-Powered Devil's Advocate. In *Proceedings of the 2024 ACM Conference on Intelligent User Interfaces*.

Chong, L.; Zhang, G.; Goucher-Lambert, K.; Kotovsky, K.; and Cagan, J. 2022. Human confidence in artificial intelligence and in themselves: The evolution and impact of confidence on adoption of AI advice. *Computers in Human Behavior*, 127: 107018.

Hemmer, P.; Schemmer, M.; Vössing, M.; and Kühl, N. 2021. Human-AI Complementarity in Hybrid Intelligence Systems: A Structured Literature Review. *PACIS*, 78.

Lu, Z.; and Yin, M. 2021. Human Reliance on Machine Learning Models When Performance Feedback is Limited: Heuristics and Risks. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–16.

Ma, S.; Lei, Y.; Wang, X.; Zheng, C.; Shi, C.; Yin, M.; and Ma, X. 2023. Who Should I Trust: AI or Myself? Leveraging Human and AI Correctness Likelihood to Promote Appropriate Trust in AI-Assisted Decision-Making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–19.

Mahmood, S. H. A.; Lu, Z.; and Yin, M. 2023. Give Weight to Human Reactions: Optimizing Complementary AI in Practical Human-AI Teams. *ICML Workshop on Artificial Intelligence and Human-Computer Interaction*.

Rechkemmer, A.; and Yin, M. 2022. When Confidence Meets Accuracy: Exploring the Effects of Multiple Performance Indicators on Trust in Machine Learning Models. In *CHI Conference on Human Factors in Computing Systems*, 1–14.

Tan, S.; Adebayo, J.; Inkpen, K.; and Kamar, E. 2018. Investigating human+ machine complementarity for recidivism predictions. *arXiv preprint arXiv:1808.09123*.

Wang, X.; Lu, Z.; and Yin, M. 2022. Will You Accept the AI Recommendation? Predicting Human Behavior in AI-Assisted Decision Making. In *Proceedings of the ACM Web Conference 2022*, 1697–1708.

Yin, M.; Wortman Vaughan, J.; and Wallach, H. 2019. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12.